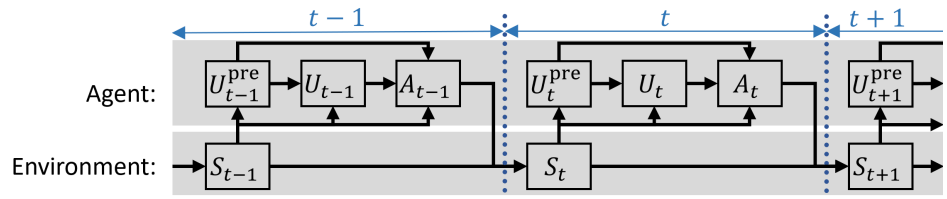

Figure 4: Diagram of the MDP and parameterized policy. This diagram can be viewed as a Bayesian network where each square is the node corresponding to the listed random variable. Bayesian network depicting the relationships of relevant random variables. Independence properties can be established by d-separation. Note that these causal properties only apply to the MDP M; any such properties of CoMDPs are explicitly proven.

A Conjugate Markov Decision Process (CoMDP)

In order to reason about the local policy gradient, we begin by modeling the i^{th} coagent's environment as an MDP, called the CoMDP, and begin by formally defining the i^{th} CoMDP. Given M, i, π_i^{pre}, π_i^{post}, and $\bar{\theta}_i$, we define a corresponding CoMDP, M^i, as $M^i \coloneqq (\mathcal{X}^i, \mathcal{U}^i, \mathcal{R}^i, P^i, R^i, d_0^i, \gamma_i)$, where:

- We write $\tilde{X}_t^i$, $\tilde{U}_t^i$, and $\tilde{R}_t^i$ to denote the state, action, and reward of M^i at time t. Below, we relate these random variables to the corresponding random variables in M. Note that all *random variables* in the CoMDP are written with tildes to provide a visual distinction between terms from the CoMDP and original MDP. Additionally, when it is clear that we are referring to the i^{th} CoMDP, we often make i implicit and denote these as $\tilde{X}_t$, $\tilde{U}_t$, and $\tilde{R}_t$.
- $\mathcal{X}^i \coloneqq \mathcal{S} \times \mathcal{U}_i^{\text{pre}}$. We often denote $\mathcal{X}^i$ simply as $\mathcal{X}$. This is the input (analogous to a state set) to the i^{th} coagent. Additionally, for $x \in \mathcal{X}$, we denote the $\mathcal{S}$ component as $x.s$ and the $\mathcal{U}^{\text{pre}}$ component as $x.u_{\text{pre}}$. We also sometimes denote an $x \in \mathcal{X}^i$ as $(x.s, x.u_{\text{pre}})$. For example, $\Pr(\tilde{X}_t^i{=}(s, u_{\text{pre}}))$ represents the probability that $\tilde{X}_t$ has $\mathcal{S}$ component s and $\mathcal{U}^{\text{pre}}$ component u_{pre}.
- $\mathcal{U}^i$ (or simply $\mathcal{U}$) is an arbitrary set that denotes the output of the i^{th} coagent.
- $\mathcal{R}^i \coloneqq \mathcal{R}$ and $\gamma_i \coloneqq \gamma$.
- $\forall x \in \mathcal{X}\ \forall x' \in \mathcal{X}\ \forall u \in \mathcal{U}\ \forall \bar{\theta}_i \in \mathbb{R}^{n-n_i}$,

$$P^i(x, u, x', \bar{\theta}_i) \coloneqq \pi_i^{\text{pre}}(x'.s, x'.u_{\text{pre}}) \sum_{a \in \mathcal{A}} P(x.s, a, x'.s) \pi_i^{\text{post}}(x, u, a),$$

  Below, we make $\bar{\theta}_i$ implicit and denote this as $P^i(x, u, x')$. Recall from the definition of an MDP and its relation to the transition function that this means: $P^i(x, u, x') = \Pr(\tilde{X}_{t+1}{=}x' | \tilde{X}_t{=}x, \tilde{U}_t{=}u)$.
- $\forall x \in \mathcal{X}\ \forall x' \in \mathcal{X}\ \forall u \in \mathcal{U}\ \forall r \in \mathcal{R}^i\ \forall \bar{\theta}_i \in \mathbb{R}^{n-n_i}$,

$$R^i(x, u, x', r, \bar{\theta}_i) \coloneqq \sum_{a \in \mathcal{A}} R(x.s, a, x'.s, r) \frac{P(x.s, a, x'.s) \pi_i^{\text{post}}(x, u, a)}{\sum_{\hat{a} \in \mathcal{A}} P(x.s, \hat{a}, x'.s) \pi_i^{\text{post}}(x, u, \hat{a})}.$$

  Like the transition function, we make $\bar{\theta}_i$ implicit and write $R^i(x, u, x', r)$.
- $\forall x \in \mathcal{X},\ d_0^i(x) \coloneqq d_0(x.s) \pi_i^{\text{pre}}(x.s, x.u_{\text{pre}})$.

We write $J_i(\theta_i)$ to denote the objective function of M^i. Notice that although $\bar{\theta}_i$ (the parameters of the other coagents) is not an explicit parameter of the objective function, it is implicitly included via the CoMDP's transition function. Note that we cannot assume that, for all θ_i, $\Delta_i(\theta_i)$ (the local policy gradient) is equivalent to $\partial J_i(\theta_i)/\partial \theta_i$ (the policy gradient of the i^{th} CoMDP); we do later prove this equivalence.

B Complete CPGT Proofs

We assume that, given the same parameters θ_i, the i^{th} coagent has the same policy in both the original MDP and the i^{th} CoMDP. That is,

Assumption 1. $\forall s \in \mathcal{S}\ \forall u_{\text{pre}} \in \mathcal{U}^{\text{pre}}\ \forall u \in \mathcal{U}\ \forall \theta_i \in \mathbb{R}^i,\ \pi_i((s, u_{\text{pre}}), u, \theta_i) = \Pr(\tilde{U}_t = u | \tilde{X}_t = (s, u_{\text{pre}}), \theta_i)$.

Property 1.

$$\forall x \in \mathcal{X},\ d_0^i(x) = \Pr(S_0 = x.s, U_0^{\text{pre}} = x.u_{\text{pre}}).$$

Proof.

$$\begin{aligned} d_0^i(x) =& d_0(x.s)\pi_i^{\text{pre}}(x.s, x.u_{\text{pre}}) \\ \overset{\text{(a)}}{=}& \Pr(S_0 = x.s) \Pr(U_0^{\text{pre}} = x.u_{\text{pre}} | S_0 = x.s) \\ =& \Pr(S_0 = x.s, U_0^{\text{pre}} = x.u_{\text{pre}}), \end{aligned}$$

where **(a)** follows from the definitions of π_i^{pre} and d_0. □

Property 2.

$$\forall s \in \mathcal{S},\ \Pr(\tilde{X}_0.s = s) = d_0(s).$$

Proof.

$$\begin{aligned} \Pr(\tilde{X}_0.s = s) \overset{\text{(a)}}{=}& \sum_{u_{\text{pre}} \in \mathcal{U}^{\text{pre}}} \Pr(\tilde{X}_0.s = s, \tilde{X}_0.u_{\text{pre}} = u_{\text{pre}}) \\ \overset{\text{(b)}}{=}& \sum_{u_{\text{pre}} \in \mathcal{U}^{\text{pre}}} d_0^i((s, u_{\text{pre}})) \\ \overset{\text{(c)}}{=}& \sum_{u_{\text{pre}} \in \mathcal{U}^{\text{pre}}} d_0(s)\pi_i^{\text{pre}}(s, u_{\text{pre}}) \\ =& d_0(s) \underbrace{\sum_{u_{\text{pre}} \in \mathcal{U}^{\text{pre}}} \Pr(U_t^{\text{pre}} = u_{\text{pre}} | S_t = s)}_{=1} \\ =& d_0(s), \end{aligned}$$

where **(a)** follows from marginalization over u_{pre}, **(b)** follows from the definition of the initial state distribution for an MDP, and **(c)** follows from the definition of d_0^i for the CoMDP (see Property 1). □

Property 3.

$$\forall x \in \mathcal{X}\ \forall x' \in \mathcal{X}\ \forall u \in \mathcal{U},\ P^i(x, u, x') = \Pr(S_{t+1} = x'.s, U_{t+1}^{\text{pre}} = x'.u_{\text{pre}} | S_t = x.s, U_t^{\text{pre}} = x.u_{\text{pre}}, U_t = u).$$

Proof.

$$\begin{aligned} P^i(x, u, x') =& \pi_i^{\text{pre}}(x'.s, x'.u_{\text{pre}}) \sum_{a \in \mathcal{A}} P(x.s, a, x'.s)\pi_i^{\text{post}}(x, u, a) \\ \overset{\text{(a)}}{=}& \sum_{a \in \mathcal{A}} \pi_i^{\text{post}}(x, u, a) \Pr(S_{t+1} = x'.s | S_t = x.s, A_t = a) \Pr(U_{t+1}^{\text{pre}} = x'.u_{\text{pre}} | S_{t+1} = x'.s) \\ \overset{\text{(b)}}{=}& \sum_{a \in \mathcal{A}} \pi_i^{\text{post}}(x, u, a) \Pr(S_{t+1} = x'.s | S_t = x.s, A_t = a) \\ & \times \Pr(U_{t+1}^{\text{pre}} = x'.u_{\text{pre}} | S_{t+1} = x'.s, S_t = x.s, A_t = a), \end{aligned}$$

where **(a)** follows from the definitions of π_i^{pre} and the transition function P, **(b)** follows from M's conditional independence properties, and $\times$ denotes scalar multiplication split across two lines. The definition of conditional probability allows us to combine the last two terms:

$$\begin{aligned} P^i(x, u, x') =& \sum_{a \in \mathcal{A}} \pi_i^{\text{post}}(x, u, a) \Pr(S_{t+1} = x'.s, U_{t+1}^{\text{pre}} = x'.u_{\text{pre}} | S_t = x.s, A_t = a) \\ \overset{\text{(a)}}{=}& \sum_{a \in \mathcal{A}} \Pr(A_t = a | S_t = x.s, U_t^{\text{pre}} = x.u_{\text{pre}}, U_t = u) \\ & \times \Pr(S_{t+1} = x'.s, U_{t+1}^{\text{pre}} = x'.u_{\text{pre}} | S_t = x.s, U_t^{\text{pre}} = x.u_{\text{pre}}, U_t = u, A_t = a) \\ \overset{\text{(b)}}{=}& \Pr(S_{t+1} = x'.s, U_{t+1}^{\text{pre}} = x'.u_{\text{pre}} | S_t = x.s, U_t^{\text{pre}} = x.u_{\text{pre}}, U_t = u), \end{aligned}$$

where **(a)** follows from the definition of π_i^{post} and the application of M's independence properties and **(b)** follows from marginalization over a. □

Property 4.

$$\begin{aligned}&\forall x \in \mathcal{X}\ \forall x' \in \mathcal{X}\ \forall u \in \mathcal{U}\ \forall r \in \mathcal{R},\ R^i(x,u,x',r)\\ =&\Pr(R_t = r | S_t = x.s, U_t^{\text{pre}} = x.u_{\text{pre}}, U_t = u, S_{t+1} = x'.s, U_{t+1}^{\text{pre}} = x'.u_{\text{pre}}).\end{aligned}$$

Proof.

$$\begin{aligned}R^i(x,u,x',r) &\coloneqq \sum_{a\in\mathcal{A}} R(x.s,a,x'.s,r)\frac{P(x.s,a,x'.s)\pi_i^{\text{post}}(x,u,a)}{\sum_{\hat{a}\in\mathcal{A}} P(x.s,\hat{a},x'.s)\pi_i^{\text{post}}(x,u,\hat{a})}\\ &\overset{\text{(a)}}{=} \sum_{a\in\mathcal{A}} R(x.s,a,x'.s,r)P(x.s,a,x'.s)\pi_i^{\text{post}}(x,u,a)\\ &\quad \div \Bigg[\sum_{\hat{a}\in\mathcal{A}} \Pr(S_{t+1} = x'.s | S_t = x.s, A_t = \hat{a}, U_t^{\text{pre}} = x.u_{\text{pre}}, U_t = u)\\ &\quad \times \Pr(A_t = \hat{a} | S_t = x.s, U_t^{\text{pre}} = x.u_{\text{pre}}, U_t = u)\Bigg]\\ &\overset{\text{(b)}}{=} \frac{\sum_{a\in\mathcal{A}} R(x.s,a,x'.s,r)P(x.s,a,x'.s)\pi_i^{\text{post}}(x,u,a)}{\Pr(S_{t+1} = x'.s | S_t = x.s, U_t^{\text{pre}} = x.u_{\text{pre}}, U_t = u)},\end{aligned}$$

where **(a)** follows from the definitions of terms in the denominator and M's conditional independence properties (applied to the first term in the denominator) and **(b)** follows from marginalization over $\hat{a}$. Expanding the definitions of the remaining terms, we get:

$$\begin{aligned}R^i(x,u,x',r) =&\frac{\sum_{a\in\mathcal{A}} \Pr(R_t = r | S_t = x.s, A_t = a, S_{t+1} = x'.s)\Pr(S_{t+1} = x'.s | S_t = x.s, A_t = a)}{\Pr(S_{t+1} = x'.s | S_t = x.s, U_t^{\text{pre}} = x.u_{\text{pre}}, U_t = u)}\\ &\times \Pr(A_t = a | S_t = x.s, U_t^{\text{pre}} = x.u_{\text{pre}}, U_t = u)\\ \overset{\text{(a)}}{=}&\frac{1}{\Pr(S_{t+1} = x'.s | S_t = x.s, U_t^{\text{pre}} = x.u_{\text{pre}}, U_t = u)}\\ &\times \sum_{a\in\mathcal{A}} \Pr(R_t = r | S_t = x.s, A_t = a, S_{t+1} = x'.s, U_t^{\text{pre}} = x.u_{\text{pre}}, U_t = u)\\ &\times \Pr(S_{t+1} = x'.s | S_t = x.s, A_t = a, U_t^{\text{pre}} = x.u_{\text{pre}}, U_t = u)\\ &\times \Pr(A_t = a | S_t = x.s, U_t^{\text{pre}} = x.u_{\text{pre}}, U_t = u),\end{aligned}$$

where **(a)** follows from M's conditional independence properties (applied to the $\Pr(R_t = r|...)$ and $\Pr(S_{t+1} = x'.s|...)$ terms). Rearranging and taking advantage of marginalization over a (the $\Pr(R_t = r|S_{t+1} = x'.s, ...)$ and $\Pr(S_{t+1} = x'.s|...)$ terms can be viewed as a union), we get:

$$\begin{aligned}R^i(x,u,x',r) =&\frac{\Pr(S_{t+1} = x'.s | S_t = x.s, U_t^{\text{pre}} = x.u_{\text{pre}}, U_t = u)}{\Pr(S_{t+1} = x'.s | S_t = x.s, U_t^{\text{pre}} = x.u_{\text{pre}}, U_t = u)}\\ &\times \Pr(R_t = r | S_t = x.s, S_{t+1} = x'.s, U_t^{\text{pre}} = x.u_{\text{pre}}, U_t = u)\\ =&\Pr(R_t = r | S_t = x.s, S_{t+1} = x'.s, U_t^{\text{pre}} = x.u_{\text{pre}}, U_t = u)\\ \overset{\text{(a)}}{=}&\Pr(R_t = r | S_t = x.s, U_t^{\text{pre}} = x.u_{\text{pre}}, U_t = u, S_{t+1} = x'.s, U_{t+1}^{\text{pre}} = x'.u_{\text{pre}}),\end{aligned}$$

where **(a)** follows from M's conditional independence properties. □

Property 5.

$$\forall s \in \mathcal{S}\ \forall u_{\text{pre}} \in \mathcal{U}_i^{\text{pre}},\ \Pr(\tilde{X}_t = (s, u_{\text{pre}})) = \Pr(S_t = s, U_t^{\text{pre}} = u_{\text{pre}}).$$

Proof.
We present a proof by induction:

Base Case:

$$\begin{aligned}\Pr(S_0 = s, U_0^{\text{pre}} = u_{\text{pre}}) &= \Pr(S_0 = s)\Pr(U_0^{\text{pre}} = u_{\text{pre}}|S_0 = s)\\ &= d_0(s)\pi_i^{\text{pre}}(s, u_{\text{pre}})\\ &= d_0^i((s, u_{\text{pre}}))\\ &= \Pr(\tilde{X}_0^i = (s, u_{\text{pre}})).\end{aligned}$$

Inductive Step:

$$\begin{aligned}\Pr(S_{t+1} = s', U_{t+1}^{\text{pre}} = u'_{\text{pre}}) &\overset{\text{(a)}}{=} \sum_{(s,u_{\text{pre}})\in\mathcal{X}} \Pr(S_t = s, U_t^{\text{pre}} = u_{\text{pre}})\Pr(S_{t+1} = s', U_{t+1}^{\text{pre}} = u'_{\text{pre}}|S_t = s, U_t^{\text{pre}} = u_{\text{pre}})\\ &\overset{\text{(b)}}{=} \sum_{(s,u_{\text{pre}})\in\mathcal{X}} \Pr(S_t = s, U_t^{\text{pre}} = u_{\text{pre}}) \sum_{u\in\mathcal{U}} \Pr(U_t = u|S_t = s, U_t^{\text{pre}} = u_{\text{pre}})\\ &\quad\times \Pr(S_{t+1} = s', U_{t+1}^{\text{pre}} = u'_{\text{pre}}|S_t = s, U_t^{\text{pre}} = u_{\text{pre}}, U_t = u)\\ &\overset{\text{(c)}}{=} \sum_{(s,u_{\text{pre}})\in\mathcal{X}} \Pr(\tilde{X}_t = (s, u_{\text{pre}})) \sum_{u\in\mathcal{U}} \Pr(\tilde{U}_t = u|\tilde{X}_t = (s, u_{\text{pre}}))\\ &\quad\times \Pr(S_{t+1} = s', U_{t+1}^{\text{pre}} = u'_{\text{pre}}|S_t = s, U_t^{\text{pre}} = u_{\text{pre}}, U_t = u),\end{aligned}$$

where **(a)** follows from marginalization over (s, u_{pre}), **(b)** follows from marginalization over u, and **(c)** is through application of the base case and Assumption 1. Notice that the last term is equivalent to P^i by Property 3, which is equivalent to the final term in the next step:

$$\begin{aligned}\Pr(S_{t+1} = s', U_{t+1}^{\text{pre}} = u'_{\text{pre}}) &= \sum_{(s,u_{\text{pre}})\in\mathcal{X}} \Pr(\tilde{X}_t = (s, u_{\text{pre}})) \sum_{u\in\mathcal{U}} \Pr(\tilde{U}_t = u|\tilde{X}_t = (s, u_{\text{pre}}))\\ &\quad\times \Pr(\tilde{X}_{t+1} = (s', u'_{\text{pre}})|\tilde{X}_t = (s, u_{\text{pre}}), \tilde{U}_t = u)\\ &\overset{\text{(a)}}{=} \sum_{(s,u_{\text{pre}})\in\mathcal{X}} \Pr(\tilde{X}_t = (s, u_{\text{pre}}))\Pr(\tilde{X}_{t+1} = (s', u'_{\text{pre}})|\tilde{X}_t = (s, u_{\text{pre}}))\\ &\overset{\text{(b)}}{=} \Pr(\tilde{X}_{t+1} = (s', u'_{\text{pre}})),\end{aligned}$$

where **(a)** and **(b)** follow from marginalization over u and (s, u_{pre}), respectively. □

Property 6.

$$\forall s \in \mathcal{S}, \Pr(S_t = s) = \Pr(\tilde{X}_t.s = s).$$

Proof.

$$\begin{aligned}\Pr(S_t = s) &= \sum_{u_{\text{pre}}\in\mathcal{U}_i^{\text{pre}}} \Pr(S_t = s, U_t^{\text{pre}} = u_{\text{pre}})\\ &\overset{\text{(a)}}{=} \sum_{u_{\text{pre}}\in\mathcal{U}_i^{\text{pre}}} \Pr(\tilde{X}_t = (s, u_{\text{pre}}))\\ &\overset{\text{(b)}}{=} \Pr(\tilde{X}_t.s = s),\end{aligned}$$

where **(a)** follows from Property 5 and **(b)** follows from marginalization over u_{pre}. □

Property 7. $\forall s \in \mathcal{S}\ \forall u_{\text{pre}} \in \mathcal{U}_i^{\text{pre}}$, $\pi_i^{\text{pre}}(s, u_{\text{pre}}) = \Pr(\tilde{X}_t.u_{\text{pre}} = u_{\text{pre}}|\tilde{X}_t.s = s)$.

Recall that $\pi_i^{\text{pre}}(s, u_{\text{pre}}) \coloneqq \Pr(U_t^{\text{pre}} = u_{\text{pre}}|S_t = s)$.

Proof.

$$\begin{aligned}
\pi_i^{\text{pre}}(s, u_{\text{pre}}) &= \Pr(U_t^{\text{pre}} = u_{\text{pre}} | S_t = s) \\
&= \frac{\Pr(U_t^{\text{pre}} = u_{\text{pre}}, S_t = s)}{\Pr(S_t = s)} \\
&\overset{\text{(a)}}{=} \frac{\Pr(\tilde{X}_t.u_{\text{pre}} = u_{\text{pre}}, \tilde{X}_t.s = s)}{\Pr(\tilde{X}_t.s = s)} \\
&= \Pr(\tilde{X}_t.u_{\text{pre}} = u_{\text{pre}} | \tilde{X}_t.s = s),
\end{aligned}$$

where **(a)** follows from properties 5 and 6. □

Property 8.

$$\begin{aligned}
&\forall s \in \mathcal{S}\ \forall s' \in \mathcal{S}\ \forall u_{\text{pre}} \in \mathcal{U}_i^{\text{pre}}\ \forall u \in \mathcal{U}, \\
&\Pr(\tilde{X}_{t+1}.s = s' | \tilde{X}_t.s = s, \tilde{X}_t.u_{\text{pre}} = u_{\text{pre}}, \tilde{U}_t = u) = \Pr(S_{t+1} = s' | S_t = s, U_t^{\text{pre}} = u_{\text{pre}}, U_t = u).
\end{aligned}$$

Proof.

$$\begin{aligned}
&\Pr(S_{t+1} = s' | S_t = s, U_t^{\text{pre}} = u_{\text{pre}}, U_t = u) \\
&\overset{\text{(a)}}{=} \sum_{a \in \mathcal{A}} \pi_i^{\text{post}}((s, u_{\text{pre}}), u, a) \Pr(S_{t+1} = s' | S_t = s, U_t^{\text{pre}} = u_{\text{pre}}, U_t = u, A_t = a) \\
&\overset{\text{(b)}}{=} \sum_{a \in \mathcal{A}} \pi_i^{\text{post}}((s, u_{\text{pre}}), u, a) \sum_{u'_{\text{pre}} \in \mathcal{U}^{\text{pre}}} \Pr(U_{t+1}^{\text{pre}} = u'_{\text{pre}} | S_t = s, U_t^{\text{pre}} = u_{\text{pre}}, U_t = u, A_t = a) \\
&\quad \times \Pr(S_{t+1} = s' | S_t = s, U_t^{\text{pre}} = u_{\text{pre}}, U_t = u, A_t = a, U_{t+1}^{\text{pre}} = u'_{\text{pre}}) \\
&\overset{\text{(c)}}{=} \sum_{a \in \mathcal{A}} \pi_i^{\text{post}}((s, u_{\text{pre}}), u, a) \sum_{u'_{\text{pre}} \in \mathcal{U}^{\text{pre}}} \Pr(U_{t+1}^{\text{pre}} = u'_{\text{pre}} | S_t = s, U_t^{\text{pre}} = u_{\text{pre}}, U_t = u, A_t = a, S_{t+1} = s') \\
&\quad \times \Pr(S_{t+1} = s' | S_t = s, U_t^{\text{pre}} = u_{\text{pre}}, U_t = u, A_t = a) \\
&\overset{\text{(d)}}{=} \sum_{a \in \mathcal{A}} \pi_i^{\text{post}}((s, u_{\text{pre}}), u, a) \sum_{u'_{\text{pre}} \in \mathcal{U}^{\text{pre}}} \Pr(U_{t+1}^{\text{pre}} = u'_{\text{pre}} | S_{t+1} = s') \Pr(S_{t+1} = s' | S_t = s, A_t = a),
\end{aligned}$$

where **(a)** follows from marginalization over a and the definition of π_i^{post}, **(b)** follows from marginalization over u'_{pre}, **(c)** follows from the fact that (abbreviating and leaving out the common givens) $\Pr(u'_{\text{pre}}) \Pr(s' | u'_{\text{pre}}) = \Pr(u'_{\text{pre}} | s') \Pr(s')$, and **(d)** follows from M's conditional independence properties (applied to the second and third terms). Notice that the second and third terms above are equivalent to P and π_i^{pre}; plugging those in and rearranging:

$$\begin{aligned}
\Pr(\tilde{X}_{t+1}.s = s' | \tilde{X}_t.s = s, \tilde{X}_t.u_{pre} &= \sum_{u'_{\text{pre}} \in \mathcal{U}^{\text{pre}}} \pi_i^{\text{pre}}(s', u'_{\text{pre}}) \sum_{a \in \mathcal{A}} P(s, a, s') \pi_i^{\text{post}}((s, u_{\text{pre}}), u, a) \\
&\overset{\text{(a)}}{=} \sum_{u'_{\text{pre}} \in \mathcal{U}^{\text{pre}}} P^i((s, u_{\text{pre}}), u, (s', u'_{\text{pre}})) \\
&\overset{\text{(b)}}{=} \sum_{u'_{\text{pre}} \in \mathcal{U}^{\text{pre}}} \Pr(\tilde{X}_{t+1}.u_{\text{pre}} = u'_{\text{pre}} | \tilde{X}_t = (s, u_{\text{pre}}), \tilde{U}_t = u) \\
&\quad \times \Pr(\tilde{X}_{t+1}.s = s' | \tilde{X}_{t+1}.u_{\text{pre}} = u'_{\text{pre}}, \tilde{X}_t = (s, u_{\text{pre}}), \tilde{U}_t = u) \\
&\overset{\text{(c)}}{=} \Pr(\tilde{X}_{t+1}.s = s' | \tilde{X}_t.s = s, \tilde{X}_t.u_{\text{pre}} = u_{\text{pre}}, \tilde{U}_t = u),
\end{aligned}$$

where **(a)** follows from the definition of P^i for the CoMDP, **(b)** follows from the definition of conditional probability, and **(c)** follows from marginalization over u'_{pre}. □

Property 9.

$$\begin{aligned}
&\forall s \in \mathcal{S}\ \forall s' \in \mathcal{S}\ \forall u_{\text{pre}} \in \mathcal{U}_i^{\text{pre}}\ \forall u'_{\text{pre}} \in \mathcal{U}_i^{\text{pre}}\ \forall u \in \mathcal{U}, \\
&\Pr(\tilde{X}_{t+1}.u_{\text{pre}} = u'_{\text{pre}} | \tilde{X}_{t+1}.s = s') = \Pr(\tilde{X}_{t+1}.u_{\text{pre}} = u'_{\text{pre}} | \tilde{X}_{t+1}.s = s', \tilde{X}_t = (s, u_{\text{pre}}), \tilde{U}_t = u).
\end{aligned}$$

Proof.

$$\begin{aligned}
&\Pr(\tilde{X}_{t+1}.u_{\text{pre}} = u'_{\text{pre}} | \tilde{X}_{t+1}.s = s', \tilde{X}_t = (s, u_{\text{pre}}), \tilde{U}_t = u) \\
&= \frac{\Pr(\tilde{X}_{t+1}.u_{\text{pre}} = u'_{\text{pre}}, \tilde{X}_{t+1}.s = s' | \tilde{X}_t = (s, u_{\text{pre}}), \tilde{U}_t = u)}{\Pr(\tilde{X}_{t+1}.s = s' | \tilde{X}_t = (s, u_{\text{pre}}), \tilde{U}_t = u)} \\
&\overset{\text{(a)}}{=} \frac{P^i((s, u_{\text{pre}}), u, (s', u'_{\text{pre}}))}{\Pr(S_{t+1} = s' | S_t = s, U^{\text{pre}} = u_{\text{pre}}, U_t = u)} \\
&= \frac{\pi_i^{\text{pre}}(s', u'_{\text{pre}}) \sum_{a \in \mathcal{A}} P(s, a, s') \pi_i^{\text{post}}((s, u_{\text{pre}}), u, a)}{\Pr(S_{t+1} = s' | S_t = s, U^{\text{pre}} = u_{\text{pre}}, U_t = u)},
\end{aligned}$$

where **(a)** follows from Property 8 applied to the denominator. Expanding the P term and applying M's conditional independence properties:

$$\begin{aligned}
&\Pr(\tilde{X}_{t+1}.u_{\text{pre}} = u'_{\text{pre}} | \tilde{X}_{t+1}.s = s', \tilde{X}_t = (s, u_{\text{pre}}), \tilde{U}_t = u) \\
&= \frac{\pi_i^{\text{pre}}(s', u'_{\text{pre}}) \sum_{a \in \mathcal{A}} \Pr(S_{t+1} = s' | S_t = s, U_t^{\text{pre}} = u_{\text{pre}}, U_t = u, A_t = a) \pi_i^{\text{post}}((s, u_{\text{pre}}), u, a)}{\Pr(S_{t+1} = s' | S_t = s, U^{\text{pre}} = u_{\text{pre}}, U_t = u)} \\
&\overset{\text{(a)}}{=} \frac{\pi_i^{\text{pre}}(s', u'_{\text{pre}}) \Pr(S_{t+1} = s' | S_t = s, U_t^{\text{pre}} = u_{\text{pre}}, U_t = u)}{\Pr(S_{t+1} = s' | S_t = s, U^{\text{pre}} = u_{\text{pre}}, U_t = u)} \\
&= \Pr(\tilde{X}_{t+1}.u_{\text{pre}} = u'_{\text{pre}} | \tilde{X}_{t+1}.s = s'),
\end{aligned}$$

where **(a)** follows from marginalization over a. □

Property 10.

$$\forall r \in \mathcal{R}, \Pr(R_t = r) = \Pr(\tilde{R}_t^i = r).$$

Proof.

$$\begin{aligned}
\Pr(R_t = r) = &\sum_{s \in \mathcal{S}} \Pr(S_t = s) \sum_{u_{\text{pre}} \in \mathcal{U}^{\text{pre}}} \Pr(U_t^{\text{pre}} = u_{\text{pre}} | S_t = s) \sum_{u \in \mathcal{U}} \Pr(U_t = u | S_t = s, U_t^{\text{pre}} = u_{\text{pre}}) \\
&\times \sum_{s' \in S} \Pr(S_{t+1} = s' | S_t = s, U_t^{\text{pre}} = u_{\text{pre}}, U_t = u) \\
&\times \sum_{u'_{\text{pre}} \in \mathcal{U}^{\text{pre}}} \Pr(U_{t+1}^{\text{pre}} = u'_{\text{pre}} | S_t = s, U_t^{\text{pre}} = u_{\text{pre}}, U_t = u, S_{t+1} = s') \\
&\times \Pr(R_t = r | S_t = s, U_t^{\text{pre}} = u_{\text{pre}}, U_t = u, S_{t+1} = s', U_{t+1}^{\text{pre}} = u'_{\text{pre}}),
\end{aligned}$$

by repeated marginalization. Applying M's conditional independence properties to the $\Pr(U_{t+1}^{\text{pre}}...)$ term:

$$\begin{aligned}
\Pr(R_t = r) = &\sum_{s \in \mathcal{S}} \Pr(S_t = s) \sum_{u_{\text{pre}} \in \mathcal{U}^{\text{pre}}} \Pr(U_t^{\text{pre}} = u_{\text{pre}} | S_t = s) \sum_{u \in \mathcal{U}} \Pr(U_t = u | S_t = s, U_t^{\text{pre}} = u_{\text{pre}}) \\
&\times \sum_{s' \in S} \Pr(S_{t+1} = s' | S_t = s, U_t^{\text{pre}} = u_{\text{pre}}, U_t = u) \sum_{u'_{\text{pre}} \in \mathcal{U}^{\text{pre}}} \Pr(U_{t+1}^{\text{pre}} = u'_{\text{pre}} | S_{t+1} = s') \\
&\times \Pr(R_t = r | S_t = s, U_t^{\text{pre}} = u_{\text{pre}}, U_t = u, S_{t+1} = s', U_{t+1}^{\text{pre}} = u'_{\text{pre}}) \\
\overset{\text{(a)}}{=} &\sum_{s \in \mathcal{S}} \Pr(\tilde{X}_t.s = s) \sum_{u_{\text{pre}} \in \mathcal{U}^{\text{pre}}} \Pr(\tilde{X}_t.u_{\text{pre}} = u_{\text{pre}} | \tilde{X}_t.s = s) \sum_{u \in \mathcal{U}} \Pr(\tilde{U}_t = u | \tilde{X}_t = (s, u_{\text{pre}})) \\
&\times \sum_{s' \in S} \Pr(\tilde{X}_{t+1}.s = s' | \tilde{X}_t = (s, u_{\text{pre}}), \tilde{U}_t = u) \sum_{u'_{\text{pre}} \in \mathcal{U}^{\text{pre}}} \Pr(\tilde{X}_{t+1}.u_{\text{pre}} = u'_{\text{pre}} | \tilde{X}_{t+1}.s = s') \\
&\times \Pr(\tilde{R}_t^i = r | \tilde{X}_t = (s, u_{\text{pre}}), \tilde{U}_t = u, \tilde{X}_{t+1} = (s', u'_{\text{pre}})),
\end{aligned}$$

where **(a)** follows from properties that show various equivalences between the two MDP's. Specifically: Property 6 (first term), Property 7 (second and fifth terms), Assumption 1 (third term), Property 8 (fourth term), and Property

4 (last term). Next, we apply Property 9 to the fifth term:

$$\begin{aligned}\Pr(R_t = r) =& \sum_{s\in\mathcal{S}} \Pr(\tilde{X}_t.s = s) \sum_{u_{\text{pre}}\in\mathcal{U}^{\text{pre}}} \Pr(\tilde{X}_t.u_{\text{pre}} = u_{\text{pre}}|\tilde{X}_t.s = s) \sum_{u\in\mathcal{U}} \Pr(\tilde{U}_t = u|\tilde{X}_t = (s, u_{\text{pre}}))\\ &\times \sum_{s'\in S} \Pr(\tilde{X}_{t+1}.s = s'|\tilde{X}_t = (s, u_{\text{pre}}), \tilde{U}_t = u)\\ &\times \sum_{u'_{\text{pre}}\in\mathcal{U}^{\text{pre}}} \Pr(\tilde{X}_{t+1}.u_{\text{pre}} = u'_{\text{pre}}|\tilde{X}_{t+1}.s = s', \tilde{X}_t = (s, u_{\text{pre}}), \tilde{U}_t = u)\\ &\times \Pr(\tilde{R}^i_t = r|\tilde{X}_t = (s, u_{\text{pre}}), \tilde{U}_t = u, \tilde{X}_{t+1} = (s', u'_{\text{pre}}))\\ \overset{\text{(a)}}{=}& (1)(1)(1)(1)(1)\Pr(\tilde{R}^i_t = r)\\ =& \Pr(\tilde{R}^i_t = r),\end{aligned}$$

where **(a)** follows from repeated marginalization. □

Lemma 1. *M^i is a Markov decision process.*

Proof. Having defined $\mathcal{X}^i$ as the state set, $\mathcal{U}^i$ as the action set, $\mathcal{R}^i$ as the reward set, P^i as the transition function, R^i as the reward function, d^i_0 as the initial state distribution, and γ_i as the discount parameter, all that remains is to ensure that P^i, R^i, and d^i_0 satisfy their necessary requirements. That is, we must show that these functions are always non-negative and that $\forall x \in \mathcal{X}, \forall u \in \mathcal{U}, \sum_{x'\in\mathcal{X}} P^i(x, u, x') = 1$, $\forall x \in \mathcal{X}, \forall u \in \mathcal{U}, \forall x' \in \mathcal{X}, \sum_{r\in\mathcal{R}^i} \mathcal{R}^i(x, u, x', r) = 1$, and $\sum_{x\in\mathcal{X}} d^i_0(x) = 1$.

The functions are always non-negative because each term in each definition is always non-negative. Next, we show that the sum over the transition function is 1:

$$\begin{aligned}&\forall x \in \mathcal{X}, \forall u \in \mathcal{U},\\ \sum_{x'\in\mathcal{X}} P^i(x, u, x') &\overset{\text{(a)}}{=} \sum_{x'\in\mathcal{X}} \Pr(S_{t+1} = x'.s, U^{\text{pre}}_{t+1} = x'.u_{\text{pre}}|S_t = x.s, U^{\text{pre}}_t = x.u_{\text{pre}}, U_t = u)\\ &= \sum_{x'.s\in\mathcal{S}} \sum_{x'.u_{\text{pre}}\in U^{\text{pre}}} \Pr(S_{t+1} = x'.s, U^{\text{pre}}_{t+1} = x'.u_{\text{pre}}|S_t = x.s, U^{\text{pre}}_t = x.u_{\text{pre}}, U_t = u)\\ &= 1,\end{aligned}$$

where **(a)** follows from Property 3. Next, we show that the sum over the reward function is 1:

$$\begin{aligned}&\forall x \in \mathcal{X},\ \forall u \in \mathcal{U}, \forall x' \in \mathcal{X},\\ \sum_{r\in\mathcal{R}^i} R^i(x, u, x', r) &\overset{\text{(a)}}{=} \sum_{r\in\mathcal{R}} \Pr(R_t = r|S_t = x.s, U^{\text{pre}}_t = x.u_{\text{pre}}, U_t = u, S_{t+1} = x'.s, U^{\text{pre}}_{t+1} = x'.u_{\text{pre}})\\ &= 1,\end{aligned}$$

where **(a)** follows from the fact that $\mathcal{R}^i := \mathcal{R}$ and from Property 4.

Finally, we show that the sum of the initial state distribution is 1:

$$\begin{aligned}\sum_{x\in\mathcal{X}} d^i_0(x) &\overset{\text{(a)}}{=} \sum_{x\in\mathcal{X}} d_0(x.s) \underbrace{\pi^{\text{pre}}_i(x.s, x.u_{\text{pre}})}_{=\Pr(U^{\text{pre}}_0 = x.u_{\text{pre}}|S_0 = x.s)}\\ &= \sum_{x.s\in\mathcal{S}} \sum_{x.u_{\text{pre}}\in U^{\text{pre}}} \Pr(S_0 = x.s, U^{\text{pre}}_0 = x.u_{\text{pre}})\\ &= 1,\end{aligned}$$

where **(a)** follows from the definition of d^i_0 for the CoMDP.

Therefore, M^i is a Markov decision process. □

Lemma 2. *For all $M, i, \pi^{\text{pre}}_i, \pi^{\text{post}}_i$, and $\bar{\theta}_i$, and given a policy parameterized by θ_i, the corresponding CoMDP M^i satisfies:*

- $\forall x \in \mathcal{X}\ \forall x' \in \mathcal{X}\ \forall u \in \mathcal{U}\ \forall r \in \mathcal{R},\ P^i(x, u, x')$
  $= \Pr(S_{t+1} = x'.s, U^{\text{pre}}_{t+1} = x'.u_{\text{pre}}|S_t = x.s, U^{\text{pre}}_t = x.u_{\text{pre}}, U_t = u)$.

- $\forall x \in \mathcal{X}\ \forall x' \in \mathcal{X}\ \forall u \in \mathcal{U}\ \forall r \in \mathcal{R}$,
  $R^i(x,u,x',r) = \Pr(R_t{=}r | S_t{=}x.s, U_t^{\text{pre}}{=}x.u_{\text{pre}}, U_t{=}u, S_{t+1}{=}x'.s, U_{t+1}^{\text{pre}}{=}x'.u_{\text{pre}})$.
- $\forall s \in \mathcal{S}\ \forall u_{\text{pre}} \in \mathcal{U}^{\text{pre}},\ \Pr(S_t = s, U_t^{\text{pre}} = u_{pre}) = \Pr(\tilde{X}_t = (s, u_{pre}))$.
- $\forall s \in \mathcal{S},\ \Pr(S_t = s) = \Pr(\tilde{X}_t.s = s)$.
- $\forall r \in \mathcal{R},\ \Pr(R_t = r) = \Pr(\tilde{R}_t^i = r)$.

Proof. This follows immediately from properties 3, 4, 5, 6, and 10. □

Property 11. *For all coagents i, for all θ_i, given the same $\theta = (\theta_i, \bar{\theta}_i)$, $J(\theta) = J_i(\theta_i)$.*

Proof.

$$\begin{aligned} J(\theta) &= \mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t | \theta\right] \\ &= \mathbf{E}\left[\sum_{t=0}^{\infty} \gamma_i^t R_t^i | \theta_i, \bar{\theta}_i,\right] \\ &= J_i(\theta_i), \end{aligned}$$

where the second step follows directly from Property 10 and the definition of γ_i. □

Lemma 3. *For all coagents i, for all θ_i, $\frac{\partial J_i(\theta_i)}{\partial \theta_i} = \Delta_i(\theta_i)$.*

Proof. In Lemma 1, we proved that the i^{th} CoMDP is an MDP. In Lemma 2, we proved that the i^{th} CoMDP correctly models the i^{th} coagent's environment. Lemma 3 follows directly from these results and the fact that Δ_i is the policy gradient for M^i (Sutton, 2000). □

Theorem 1.
$\nabla J(\theta) = [\Delta_1(\theta_1)^\intercal, \Delta_2(\theta_2)^\intercal, \ldots, \Delta_m(\theta_m)^\intercal]^\intercal$, where m is the number of coagents, and Δ_i is the local policy gradient of the ith coagent.

Proof.

$$\begin{aligned} \nabla J(\theta) &= \left[\frac{\partial J(\theta)}{\partial \theta_1}^\intercal, \frac{\partial J(\theta)}{\partial \theta_2}^\intercal, \ldots, \frac{\partial J(\theta)}{\partial \theta_m}^\intercal\right]^\intercal \\ &\overset{\text{(a)}}{=} \left[\frac{\partial J_1(\theta_1)}{\partial \theta_1}^\intercal, \frac{\partial J_2(\theta_2)}{\partial \theta_2}^\intercal, \ldots, \frac{\partial J_m(\theta_m)}{\partial \theta_m}^\intercal\right]^\intercal \\ &\overset{\text{(b)}}{=} \left[\frac{\Delta_1(\theta_1)}{\partial \theta_1}^\intercal, \frac{\Delta_2(\theta_2)}{\partial \theta_2}^\intercal, \ldots, \frac{\Delta_m(\theta_m)}{\partial \theta_m}^\intercal\right]^\intercal, \end{aligned}$$

where **(a)** follows directly from Property 11 and where **(b)** follows directly from Lemma 3. □

Corollary 1. *If α_t is a deterministic positive stepsize, $\sum_{t=0}^{\infty} \alpha_t = \infty$, $\sum_{t=0}^{\infty} \alpha_t^2 < \infty$, additional technical assumptions are met (Bertsekas & Tsitsiklis, 2000, Proposition 3), and each coagent updates its parameters, θ_i, with an unbiased local policy gradient update $\theta_i \leftarrow \theta_i + \alpha_t \widehat{\Delta}_i(\theta_i)$, then $J(\theta)$ converges to a finite value and $\lim_{t\to\infty} \nabla J(\theta) = 0$.*

Proof. Corollary 1 follows directly from the CPGT, Proposition 3 from Bertsekas & Tsitsiklis (2000), and the assumption that the discounted sum of rewards over an episode is finite (this last assumption prevents $J(\theta)$ from diverging to ∞). □

C Asynchronous Coagent Networks: Supplementary Proofs

C.1 Synchronous Network Correctness

Our goal is to show that the synchronous, acyclic reduction of our original asynchronous, cyclic network behaves identically to our original network. That is, for all $s \in \mathcal{S}, u \in \mathcal{U}^{\text{all}}, a \in \mathcal{A}, e \in \{0,1\}^m$, $\grave{\pi}((s,u),(a,u',e)) = \Pr(A_t = a, U_t^{\text{all}} = u', E_t = e | S_t = s, U_{t-1}^{\text{all}} = u)$. Because of the large number of variables, if we use one of these lowercase symbols in an equation, assume that it holds for all values in its respective set.

Proof. We present a proof by induction. We assume a topological ordering of the coagents, such that for any $j < i$, the j^{th} coagent executes before the i^{th} coagent. We perform induction over i, with the inductive assumption that the outputs of all the previous coagents, as well as their decisions whether or not to execute, correspond to the original network. The inductive hypothesis is that for all $j < i$:

$$\Pr(\grave{A}_t.u_j^{\text{all}} = u_j', \grave{A}_t.e_j = e_j | \grave{S}_t = (s,u)) = \Pr(U_t^{\text{all}}.u_j = u_j', E_t^j = e_j | S_t = s, U_{t-1}^{\text{all}} = u). \tag{1}$$

Consider the base case, $i = 1$. $\grave{\mathcal{U}}_1^{\text{pre}}$ and $\mathcal{U}_1^{\text{pre}}$ are both the empty set, because no coagents produce an output before the first coagent in either network. As a result, the distribution over the execution probability is trivially the same in both networks, that is, $\Pr(E_t^1 = 1 | S_t = s, U_{t-1}^{\text{all}} = u) = \beta_1((s,\varnothing,u)) = \Pr(\grave{A}_t.e_1 = 1 | \grave{S}_t = (s,u))$. Next, we consider the action. If the coagent executes, $\Pr(U_t^1 = u_1' | E_t^1 = 1, S_t = s, U_{t-1}^{\text{all}} = u) = \pi_1((s,\varnothing,u),u_1') = \Pr(\grave{A}_t.u_1^{\text{all}} = u_1' | A_t.e_1 = 1, \grave{S}_t = (s,u))$. If the coagent does not execute, the action is trivially u_1 in both cases. Therefore, Equation (1) holds for $j = 1$.

Next we consider the inductive step, where we show that Equation (1) holds for the i^{th} coagent given that it holds for $j < i$. First we consider the execution function, the output of which is represented in the synchronous setting by $\grave{A}_t.e_i$, and in the asynchronous setting by E_t. In the asynchronous setting, the probability of the i^{th} coagent executing at time step t is $\beta_i((S_t, U_t^{\text{pre}}, U_{t-1}^{\text{all}}))$. Since we are not given U_t^{pre}, we must sum over possible values:

$$\Pr(E_t^i = 1 | S_t = s, U_t^{\text{all}} = u) = \sum_{u_{\text{pre}} \in \mathcal{U}_i^{\text{pre}}} \beta_i((s, u_{\text{pre}}, u)) \Pr(U_t^{\text{pre}} = u_{\text{pre}} | S_t = s, U_{t-1}^{\text{all}} = u).$$

In the reduced setting, we instead have a coagent, such that $\Pr(\grave{A}_t.e_i = 1 | \grave{S}_t = (s,u)) = \beta_i((s, \grave{U}_t^{\text{pre}}, u))$. Again, we sum over possible values of $\grave{U}_t^{\text{pre}}$:

$$\Pr(\grave{A}_t.e_i = 1 | \grave{S}_t = (s,u)) = \sum_{u_{\text{pre}} \in \mathcal{U}_i^{\text{pre}}} \beta_i((s, u_{\text{pre}}, u)) \Pr(\grave{U}_t^{\text{pre}} = u_{\text{pre}} | \grave{S}_t = (s,u)).$$

Recall the reduced setting was defined such that for all $j < i, \grave{A}_t.u_j^{\text{all}} = \grave{U}_t^{\text{pre}}.u_j$, and in the asynchronous setting, $U_t^{\text{pre}}.u_j = U_t^{\text{all}}.u_j$. We therefore can conclude from (1) and by substitution that for all $j < i, \Pr(\grave{U}_t^{\text{pre}}.u_j = u | \grave{S}_t = (s,u)) = \Pr(U_t^{\text{pre}}.u_j = u | S_t = s, U_{t-1}^{\text{all}} = u)$. Substituting this into the above equations:

$$\begin{aligned}\Pr(\grave{A}_t.e_i = 1 | \grave{S}_t = (s,u)) &= \sum_{u_{\text{pre}} \in \mathcal{U}_i^{\text{pre}}} \beta_i((s, u_{\text{pre}}, u)) \Pr(U_t^{\text{pre}} = u_{\text{pre}} | S_t = s, U_{t-1}^{\text{all}} = u) \\ &= \Pr(E_t^i = 1 | S_t = s, U_t^{\text{all}} = u).\end{aligned}$$

Note also that from the perspective of $\grave{\pi}_i$, $\grave{A}_t.e_i = \grave{U}_t^{\text{pre}}.e_i$. Next we consider the output of the i^{th} coagent, given in the asynchronous setting as U_t^i, and in the reduced setting by $\grave{A}_t.u_i^{\text{all}}$. In the original setting, U_t^i was given such that for all $u_i \in U^i$:

$$\Pr(U_t^i = u_i' | S_t = s, U_{t-1}^{\text{all}} = u, U_t^{\text{pre}} = u_{\text{pre}}, E_t^i = e_i) = \begin{cases} \pi_i((s, u_{\text{pre}}, u), u_i'), & \text{if } e_i = 1 \\ 1, & \text{if } e_i = 0 \text{ and } u_i' = u_i \\ 0, & \text{otherwise.} \end{cases}$$

In the synchronous setting, we are given:

$$\begin{aligned}\Pr(\grave{A}_t.u_i^{\text{all}} = u_i' | \grave{S}_t = (s,u), \grave{U}_t^{\text{pre}}.u = u_{\text{pre}}, \grave{U}_t^{\text{pre}}.e_i = e_i) &= \grave{\pi}_i(((s,u), u_{\text{pre}}), u) \\ &= \begin{cases} \pi_i((s, u_{\text{pre}}, u), u_i'), & \text{if } e_i = 1 \\ 1, & \text{if } e_i = 0 \text{ and } u_i' = u_i \\ 0, & \text{otherwise.} \end{cases}\end{aligned}$$

Since we were given s and u, assumed through the inductive hypothesis that $\Pr(\grave{U}_t^{\text{pre}}.u = u_{\text{pre}}|\grave{S}_t = (s,u)) = \Pr(U_t^{\text{pre}} = u_{\text{pre}}|S_t = s, U_{t-1}^{\text{all}} = u)$, and showed that $\Pr(\grave{U}_t^{\text{pre}}.e_i = e_i|\grave{S}_t = (s,u)) = \Pr(E_t^i = e_i|S_t = s, U_{t-1}^{\text{all}} = u)$, we know that the distributions over the variables we conditioned on are equal. Since we also showed that the conditional distributions are equal, we conclude that $\Pr(\grave{A}_t.u_i^{\text{all}} = u_i'|\grave{S}_t = (s,u)) = \Pr(U_t^i = u_i'|S_t = s, U_{t-1}^{\text{all}} = u)$.

This completes the inductive proof that $\Pr(\grave{A}_t.u^{\text{all}} = u', \grave{A}_t.e = e|\grave{S}_t = (s,u)) = \Pr(U_t^{\text{all}} = u', E_t = e|S_t = s, U_{t-1}^{\text{all}} = u)$. We still must consider $\grave{A}_t.a$. This is given by the output of some predefined subset of coagents, which is the same subset in both the synchronous and asynchronous network. We showed that the distribution over outputs was the same for corresponding coagents in the two networks, and therefore can conclude immediately that $Pr(\grave{A}_t.a = a|\grave{S}_t = (s,u)) = \Pr(A_t = a|S_t = s, U_{t-1}^{\text{all}} = u)$. Finally:

$$\begin{aligned}\Pr(A_t = a, U_t^{\text{all}} = u', E_t = e|S_t = s, U_{t-1}^{\text{all}} = u) &= \Pr(\grave{A}_t.a = a, \grave{A}_t.u^{\text{all}} = u', \grave{A}_t.e = e|\grave{S}_t = (s,u)) \\ &= \Pr(\grave{A}_t = (a,u',e)|\grave{S}_t = (s,u)) \\ &= \grave{\pi}((s,u),(a,u',e)).\end{aligned}$$

□

C.2 Equivalence of Objectives

In both settings, the network depends on the same parameter vector, θ. In this section, we show that for all settings of this parameter vector the resulting sum of rewards is equivalent in both settings. That is, $J(\theta) = \grave{J}(\theta)$.

Proof. We begin by showing that the distribution over the "true" states and actions is equal in both settings, that is, for all $s \in \mathcal{S}, a \in \mathcal{A}$, $Pr(\grave{S}_t.s = s, \grave{A}_t.a = a) = \Pr(S_t = s, A_t = a)$. Once this is shown, we show that the reward distributions are the same, that is, for all r, $\Pr(\grave{R}_t = r) = \Pr(R_t = r)$. Finally, we show $J(\theta) = \grave{J}(\theta)$.

C.2.1 Equivalence of State Distributions

First, we show that $\Pr(\grave{S}_t = (s,u)) = \Pr(S_t = s, U_{t-1}^{\text{all}} = u)$, by induction over time steps. The base case is the initial state, $\grave{S}_0$. We assumed in the problem setup for the asynchronous setting that for all i and j, the random variables S_0, U_{-1}^i, and U_{-1}^j are independent. For all s and u:

$$\begin{aligned}\Pr(\grave{S}_0 = (s,u)) &= \grave{d}_0((s,u)) \\ &= d_0(s)\prod_{i=1}^m h_0^i(u_i) \\ &= \Pr(S_0 = s)\prod_{i=1}^m \Pr(U_{-1}^i = u_i) \\ &= \Pr(S_0 = s, U_{-1}^{\text{all}} = u) \\ &= \Pr(S_0 = s, U_{-1}^{\text{all}} = u).\end{aligned}$$

Thus, we've proven the base case. Next we consider the inductive step:

$$\begin{aligned}&\Pr(\grave{S}_{t+1} = (s',u')|\grave{S}_t = (s,u)) \\ &= \sum_{(a,u'',e)\in\grave{\mathcal{A}}} \Pr(\grave{S}_{t+1} = (s',u')|\grave{S}_t = (s,u), \grave{A}_t = (a,u'',e)) \Pr(\grave{A}_t = (a,u'',e)|\grave{S}_t = (s,u)) \\ &= \sum_{(a,u'',e)\in\grave{\mathcal{A}}} \grave{P}((s,u),(a,u'',e),(s',u'))\grave{\pi}((s,u),(a,u'',e)) \\ &= \sum_{a\in\mathcal{A},u''\in\mathcal{U}^{\text{all}},e\in\mathcal{E}} \begin{cases} P(s,a,s')\grave{\pi}((s,u),(a,u'',e)) & \text{if } u' = u'' \\ 0 & \text{otherwise,}\end{cases}\end{aligned}$$

The case statement comes from the definition of $\grave{P}$. Clearly, we can eliminate all of the parts of the summation where $u' \neq u''$. Therefore:

$$\Pr(\grave{S}_{t+1} = (s',u')|\grave{S}_t = (s,u)) = \sum_{a\in\mathcal{A},e\in\mathcal{E}} P(s,a,s')\grave{\pi}((s,u),(a,u',e)).$$

Next, we can apply the equivalence shown in section C.1 and the definition of P:

$$\sum_{a\in\mathcal{A},e\in\mathcal{E}} P(s,a,s')\grave{\pi}((s,u),(a,u',e))$$
$$= \sum_{a\in\mathcal{A},e\in\mathcal{E}} \Pr(S_{t+1}=s'|A_t=a,S_t=s)\Pr(A_t=a,U_t^{\text{all}}=u',E_t=e|S_t=s,U_{t-1}^{\text{all}}=u).$$

By the law of total probability, we can eliminate E_t from the expression:

$$\sum_{a\in\mathcal{A}} \Pr(S_{t+1}=s'|A_t=a,S_t=s)\Pr(A_t=a,U_t^{\text{all}}=u'|S_t=s,U_{t-1}^{\text{all}}=u).$$

Next, S_{t+1} is conditionally independent of U_t^{all} and U_{t-1}^{all} given S_t and A_t, so we can rewrite the expression as a sum over a single probability:

$$\sum_{a\in\mathcal{A}} \Pr(S_{t+1}=s'|A_t=a,U_t^{\text{all}}=u',S_t=s,U_{t-1}^{\text{all}}=u)\Pr(A_t=a,U_t^{\text{all}}=u'|S_t=s,U_{t-1}^{\text{all}}=u)$$
$$= \sum_{a\in\mathcal{A}} \Pr(S_{t+1}=s',A_t=a,U_t^{\text{all}}=u'|S_t=s,U_{t-1}^{\text{all}}=u).$$

Finally, by the law of total probability, we eliminate A_t:

$$\sum_{a\in\mathcal{A}} \Pr(S_{t+1}=s',A_t=a,U_t^{\text{all}}=u'|S_t=s,U_{t-1}^{\text{all}}=u) = \Pr(S_{t+1}=s',U_t^{\text{all}}=u'|S_t=s,U_{t-1}^{\text{all}}=u).$$

Thus, the inductive hypothesis holds. We have therefore shown that for all t, $\Pr(\grave{S}_t=(s,u)) = \Pr(S_t=s,U_{t-1}^{\text{all}}=u)$.

C.2.2 Equivalence of Reward Distributions

It follows immediately from the above equality and C.1 that $\Pr(\grave{A}_t=(a,u,e)) = \Pr(A_t=a,U_t^{\text{all}}=u,E_t=e)$. We turn our attention to the reward distribution:

$$\begin{aligned}
&\Pr(\grave{R}_t=r)\\
&=\sum_{(s,u)\in\grave{\mathcal{S}}}\sum_{(a,u',e)\in\grave{\mathcal{A}}}\sum_{(s',u'')\in\grave{\mathcal{S}}} \Pr(\grave{R}_t=r|\grave{S}_t=(s,u),\grave{A}_t=(a,u',e),\grave{S}_{t+1}=(s',u''))\\
&\quad\times\Pr(\grave{S}_t=(s,u),\grave{A}_t=(a,u',e),\grave{S}_{t+1}=(s',u''))\\
&=\sum_{(s,u)\in\grave{\mathcal{S}}}\sum_{(a,u',e)\in\grave{\mathcal{A}}}\sum_{(s',u'')\in\grave{\mathcal{S}}} \grave{R}((s,u),(a,u',e),(s',u''))\Pr(\grave{S}_t=(s,u),\grave{A}_t=(a,u',e),\grave{S}_{t+1}=(s',u''))\\
&=\sum_{(s,u)\in\grave{\mathcal{S}}}\sum_{(a,u',e)\in\grave{\mathcal{A}}}\sum_{(s',u'')\in\grave{\mathcal{S}}} R(s,a,s')\Pr(\grave{S}_t=(s,u),\grave{A}_t=(a,u',e),\grave{S}_{t+1}=(s',u''))\\
&=\sum_{(s,u)\in\grave{\mathcal{S}}}\sum_{(a,u',e)\in\grave{\mathcal{A}}}\sum_{(s',u'')\in\grave{\mathcal{S}}} R(s,a,s')\Pr(S_t=s,U_{t-1}^{\text{all}}=u,A_t=a,U_t^{\text{all}}=u',E_t=e,S_{t+1}=s')\\
&=\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}} R(s,a,s')\Pr(S_t=s,A_t=a,S_{t+1}=s')\\
&=\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}} \Pr(R_t=r|S_t=s,A_t=a,S_{t+1}=s')\Pr(S_t=s,A_t=a,S_{t+1}=s')\\
&=\Pr(R_t=r).
\end{aligned}$$

C.2.3 Equivalence of Objectives

Finally, we show the objectives are equal:

$$\grave{J}(\theta) = \mathbf{E}[\sum_{t=0}^{T}\grave{\gamma}^t\grave{R}_t] = \mathbf{E}[\sum_{t=0}^{T}\gamma^t R_t] = J(\theta),$$

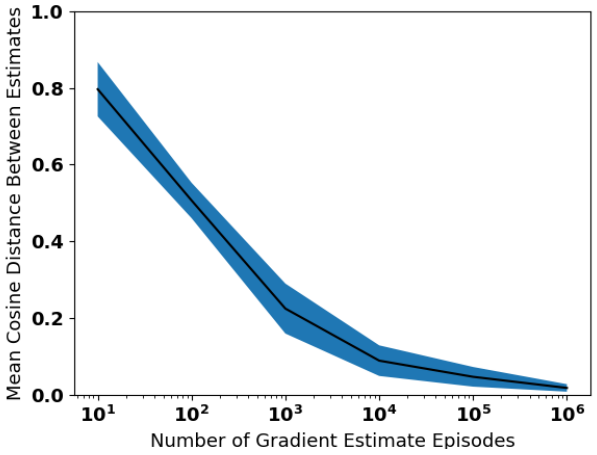

Figure 5: The average (across coagents) cosine distance between the gradient estimates of the finite difference method and the CPGT (vertical axis) versus the number of episodes used for the CPGT estimate (horizontal axis). This data is drawn from 20 trials. Error bars represent standard error. 5×10^8 episodes are used for each finite difference estimate. As the amount of data used for the CPGT gradient estimate increases, the cosine distance approaches zero, indicating that the two gradient estimates converge to the same value as the amount of data increases.

by linearity of expectation. □

D Experimental Details of Finite Difference Comparison

To empirically test the Asynchronous Coagent Policy Gradient Theorem (ACPGT), we compare the gradient (∇J) estimates of the ACPGT and a finite difference method. Finite difference methods are a well-established technique for computing the gradient of a function from samples; they serve as a straightforward baseline to evaluate the gradients produced by our algorithm. We expect these estimates to approach the same value as the amount of data used approaches infinity. For the purposes of testing the ACPGT, we use a simple toy problem and an asynchronous coagent network. The results are presented in Figure 5; this data provides empirical support for the ACPGT.

We use a simple 3×3 Gridworld. The network structure used in this experiment consists of three coagents with tabular state-action value functions and softmax policies: Two coagents receive the tabular state as input, and each of those two coagents have a single tabular binary output to the third coagent, which in turn outputs the action (up, down, left, or right). This results in two coagents with 18 parameters each, and one coagent with 16 parameters, resulting in a network with 52 parameters. The coagents asynchronously execute using a geometric distribution; the environment updates every step and each coagent has a 0.5 probability of executing each step. The gradient estimates appear to converge, providing empirical support of the CPGT. The data is drawn from 20 trials. 5×10^8 episodes were used for each finite difference estimate. For each trial, five training episodes were conducted before the parameters were frozen and the two gradient estimation methods were run. The coagents were trained with Sutton & Barto's (2018) actor-critic with eligibility traces algorithm and shared a single critic. Note that the critic played no role in the gradient estimation methods, only in the initial training episodes. Hyperparmaters used: critic step size $= 0.024686$, $\gamma = 1$, input agent step size $= 0.02842$, output agent step size $= 0.1598$, and all agents' $\lambda = 0.8085$.

Note that, while a large amount of data is required to reduce the cosine distance to near 0, this does not reflect how long the network takes to learn near-optimal behavior. Figure 6 depicts the mean episodic return of 10,000 trials of 200 episodes each (the same environment, algorithms, network structure, hyperparameters, etc. described above). Despite its handicap of only having a coagent execute with a 0.5 probability at each time step (a rather significant handicap for this network structure in a gridworld), the network achieves near-optimal returns relatively quickly.

E Option-Critic

E.1 Option-Critic Complete Description

In this section, we adhere mostly to the notation given by Bacon et al. (2017)'s, with some minor changes used to enhance conceptual clarity regarding the inputs and outputs of each policy. In the option-critic framework, the

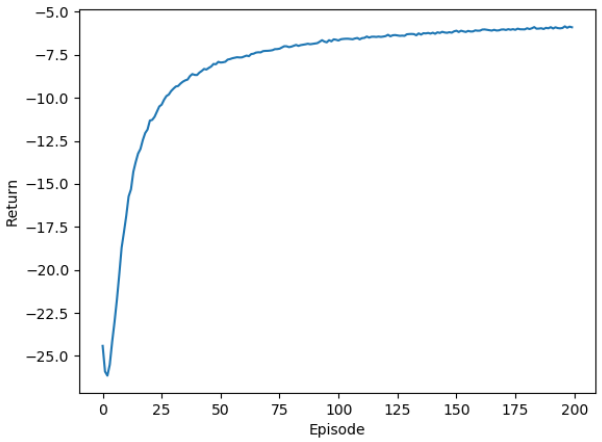

Figure 6: The mean learning curve from 10,000 trials of coagent network described in Section D (without freezing the parameters after 5 episodes). While a large amount of data is required to reduce the cosine distance to near 0, note that this fact does not reflect how long the network takes to learn near-optimal behavior.

agent is given a set of *options*, Ω. The agent selects an option, $\omega \in \Omega$, by sampling from a policy $\pi_\Omega : \mathcal{S} \times \Omega \to [0, 1]$. An action, $a \in \mathcal{A}$, is then selected from a policy which considers both the state and the current option: $\pi_\omega : (\mathcal{S} \times \Omega) \times \mathcal{A} \to [0, 1]$. A new option is not selected at every time step; rather, an option is run until a *termination function*, $\beta : (\mathcal{S} \times \Omega) \times \{0, 1\} \to [0, 1]$, selects the termination action, 0. If the action 1 is selected, then the current option continues. π_ω is parameterized by weights θ, and β by weights ϑ.

E.2 Option-Critic Gradient Equivalence

The APCGN expression gives us $\frac{\partial J}{\partial \vartheta} = \sum_{x \in (\mathcal{S} \times \Omega)} d^\pi_\Omega(x) \sum_{u \in \{0,1\}} \frac{\partial \beta(x,u)}{\partial \vartheta} Q_\beta(x, u)$. We will show that this is equivalent to Bacon et al. (2017)'s expression for $\frac{\partial J}{\partial \vartheta}$. Note that we only have two actions, whose probabilities must sum to one. Therefore, the gradients of the policy are equal in magnitude but opposite in sign. That is, for all $x \in (\mathcal{S} \times \Omega)$: $\beta(x, 0) + \beta(x, 1) = 1$, so $\frac{\partial \beta(x,0)}{\partial \vartheta} = -\frac{\partial \beta(x,1)}{\partial \vartheta}$. Additionally, we know that $Q_\beta(x, 1)$ is the expected value of continuing option $x.\omega$ in state $x.s$, given by $Q_\Omega(x.s, x.\omega)$. $Q_\beta(x, 0)$ is the expected value of choosing a new action in state $x.s$, given by $V_\Omega(x.s)$, and therefore, $Q_\beta(x, 1) - Q_\beta(x, 0) = A_\Omega(x.s, x.\omega)$. The full derivation is:

$$\begin{aligned}
\frac{\partial J}{\partial \vartheta} &= \sum_{x \in (\mathcal{S} \times \Omega)} d^\pi_\Omega(x) \left[\frac{\partial \beta(x, 0)}{\partial \vartheta} Q_\beta(x, 0) + \frac{\partial \beta(x, 1)}{\partial \vartheta} Q_\beta(x, 1) \right] \\
&= -\sum_{x \in (\mathcal{S} \times \Omega)} d^\pi_\Omega(x) \frac{\partial \beta(x, 0)}{\partial \vartheta} (Q_\beta(x, 1) - Q_\beta(x, 0)) \\
&= -\sum_{x \in (\mathcal{S} \times \Omega)} d^\pi_\Omega(x) \frac{\partial \beta(x, 0)}{\partial \vartheta} A_\Omega(x.s, x.\omega).
\end{aligned}$$

We see that the result is exactly equivalent to the expression for $\partial J / \partial \vartheta$ derived by Bacon et al. (2017).